Università degli Studi di Cagliari

DOTTORATO DI RICERCA IN

Terapia Pediatrica e Farmacologia dello Sviluppo

Ciclo XXVII

Joint whole exome sequencing and linkage analysis in a multigenerational family segregating Type 1 Diabetes

Settore scientifico disciplinare di afferenza

MED/03 – Genetica Medica

Presentata da: Elisabetta Mereu

Coordinatore Dottorato: Prof. Paolo Moi

Tutors/Relatori: Dott.ssa Serena Sanna, Prof. Paolo Moi

Esame finale anno accademico 2013 – 2014

Elisabetta Mereu gratefully acknowledges Sardinia Regional Government for the financial support of her PhD scholarship (P.O.R. Sardegna F.S.E. Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2007-2013 - Axis IV Human Resources, Objective l.3, Line of Activity l.3.1.).

Acknowledgements

This work was supported by resources from the National Research Council (CNR) of Cagliari - Istituto di Ricerca Genetica e Biomedica (IRGB). The project was conceived by the Director of the Institute Prof. Francesco Cucca, who also provided reagents and funding sources. I am thankful to the colleagues who contributed to this project, in particular the Immunogenetics, Genotyping and Sequencing Platforms (GSP) and High-Performance Computing (HPC) laboratories at CRS4 in Pula, as well as the genotyping and statistical groups at the IRGB Institute. I also thank the group of biologists at IRGB who is carrying out functional experiments on the preliminary results of this thesis.

I acknowledge the special contribution of all the volunteers participating in this research project and all the clinicians and physicians for their continuous effort in patients'enrollment. I am especially grateful to my tutors Serena Sanna and Prof. Paolo Moi for their continuous advices, and to the Director of the IRGB Institute Prof. Francesco Cucca for involving me

in this project. Finally, I also thank my family, friends and colleagues for supporting me morally in this difficult but exciting research experience.

Contents

Chapter 1

Introduction

In the past decade, genetic analyses of complex diseases were carried out using a systematic evaluation of the genome, known as genome-wide association scan (GWAS). This approach consists in an examination of many (hundreds of thousands) common genetic variants in thousands of individuals to see if any variant is associated with a trait. The underlying idea is that common diseases must be driven by common variants, theory also known as common variant - common disease hypothesis [1]. To date, GWAS had highlighted thousands of common genetic loci associated with diseases susceptibility and findings are reproducible across populations. Despite its success, this approach has also limitations. In the most cases, GWAS hits fall outside of annotated genes, limiting our understanding on their molecular function. Furthermore, these associations have often very

small effects on diseases'risk and account only for a modest portion of the estimated genetic components of many diseases, leaving a substantial fraction unknown [2]. Several hypothesis have been made on where the missing heritability is [2]. One attributes a major role to rare variants, a category of genetic changes that are present in $< 1\%$ of the population, mostly ignored by GWAS.

Today, the advent of next-generation sequencing (NGS) has revolutionized the study of genetic variations, as the entire spectrum of genomic variation can be assessed, including rare variants. Cost to sequence all exons or genomes has considerably reduced, allowing sequencing of increasingly large number of samples.

Whole-genome sequencing represents the most complete variant-discovery strategy, and a common strategy in large scale studies is to incorporate whole-genome sequencing information into GWAS with inferential methods known as genotype imputation [3]. Exome sequencing only focuses on coding regions and thus limits the spectrum of variants tested, but has been widely applied as a powerful alternative cost-effective approach to detect rare coding variants, which often have more marked functional consequences [4] [5]. Specifically, exome-sequencing technologies have been mostly important for the identification of molecular defects in patients with Mendelian disorder or as a diagnostic tool in case of suspected rare genetic disorder [6]. This success led researchers to extend it in the case of complex diseases, in order to identify rare coding variants that are not detected by

GWAS [7]. Rare variants however are hard to identify in the large scale population studies used in GWAS: their often occur too infrequently to allow association testing in a sufficient number of individuals. One strategy is to focus on cohorts that were initially collected for linkage analysis - i.e., families with multiple affected individuals ; in the assumption that affected individuals drawn from the same family must share the same causal variant, families are a natural enriched setting for very rare variants, as the same variant will be carried by multiple chromosomes. The challenge in such analyses is that even randomly selected individuals in any family will share substantial fractions of their genomes in common, therefore the prioritization of individual variants detected trough sequencing involves additional criteria. For example, a combination of exome sequencing in key individuals with family-based linkage analysis using classical genotyping arrays in the full family may increases power to detect rare genetic variants with large effect size. In fact, exome sequencing allows detection of a set of potential candidate rare variants, and linkage analysis will help to distinguish those that are disease-causing by evaluating co-segregation with the phenotype within the family. Examples of successful applications of whole-exome sequencing in family-based designs [8][9], also with a limited number of cases [10], have been already reported for complex diseases [8].

In this PhD thesis, I will present an application of a joint exome sequencing and linkage analysis in a multigenerational family with multiple mem-

bers affected by Type 1 Diabetes (T1D), a form of diabetes mellitus that results from the autoimmune destruction of the insulin-producing beta cells in the pancreas and that is usually diagnosed in children and young adults. From this family, we recruited 13 members in 3 generations, 9 of which were affected. All samples were genotyped for about 700,000 common variants using the Illumina OmniExpress array, and exome-sequencing was carried in 3 affected individuals (one per generation) and a healthy individual (pedigree is showed in Figure 1.1).

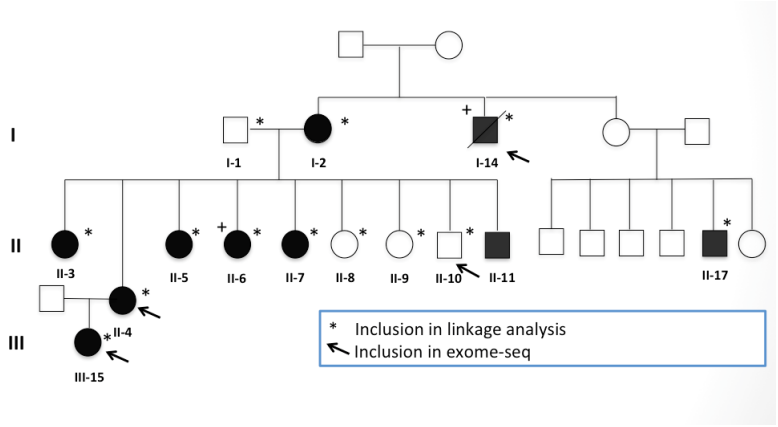


Figure 1.1: The patients family tree showing selected individuals in the exome-seq and linkage analysis.

The thesis is organized as follows:

- Chapter 2 describes the state of the art in current literature on genetics of Type 1 diabetes.

- Chapter 3 details methods and analysis protocols, including description of the samples, study design and workflow in whole-exome sequencing analysis, from the reads alignment, variant calling, including reads quality controls to the variant annotations, filtering and association analysis. I will also present data processing and quality for the genotyping arrays, which will be integrated with exome-sequencing in a parametric linkage analysis on the family.

- Chapter 4 will be dedicate to conclusions and future developments of this project.

1.1 Rare variants in complex diseases

Our understanding on the genetic architecture of complex diseases changed quite rapidly with the advent of the GWAS approach. In particular, for most of the diseases it appears that heritability follows a polygenic model, with tens or even hundreds variants modulating the diseases risk. GWAS have extensively searched for common variants with large or moderate effects, and resulting observations show that the total contribution of this category of variants to the disease does not fully explain the estimated genetic heritability [12]. In fact, as statistical power in GWAS depends by a combination of effect size and minor allele frequency, many disease susceptibility loci with either a very small effect size or highly-penetrant

rare alleles have been missed by the GWAS approach.
Several population genetic studies suggest that low and rare variants tend to be recent in origin and support their possible contribute to disease risk, being enriched for potentially functional mutations [15] [16]. Such variants are likely to be seen in families with a number of affected individuals exciding expectations based on the overall population risk. Family-based studies are thus valuable in the search of rare variants.

1.2 Next generation sequencing to detect rare variants in complex diseases

Over the past few years, scientific discoveries made through whole genome and exome sequencing grew exponentially, accelerating genetic studies on Mendelian and complex diseases and increasing the total number of known genomic variations. Large international efforts, as the *Exome Sequencing Project*[1] *(ESP)* and *1000 Genomes Project*[2] *(1000G)*, have sequenced thousands of individuals from different nationality, with the common goal to characterize the geographic and functional spectrum of human genetic variation [14]. The 1000 Genomes project includes sequencing of 2535 individuals sampled from 26 populations drawn from the five continents, analyzed through different levels of resolution. The overall num-

[1] https://esp.gs.washington.edu/drupal/
[2] http://www.1000genomes.org/

ber of variations discovered is 79 million sites, including biallelic single-nucleotide polymorphisms (SNPs), insertion, deletions (indels), complex short substitutions and other structural variant classes [14]. The NHLBI Exome Sequencing Project sequenced the coding region of 15,585 genes by whole-exome sequencing in up to 6500 individuals of African and European ancestry, finding thousands of very rare variants, some of which are population-specific [16]. Therefore, sequencing of human genomes and exomes revealed that most low and rare variants showed substantial geographic differentiation and that the majority of protein-coding variation predicted functionally important were rare [14] [16]. Characterization of such variants, likely evolutionarily recent and under a weak selection pressure [16], will allow a better interpretation of their functional role in specific studies.

1.2.1 Whole exome sequencing

Whole exome sequencing (WES) is a technique to selectively capture and sequence protein-coding regions in parallel through high-throughput sequencing. Similarly to whole-genome sequencing, the targeted DNA is sequenced by dividing it in millions of small fragments that are then sequenced in parallel concurrently. Then all fragments are assembled back together by aligning the sequenced nucleotides to a known reference genome or by algorithms for de-novo assembly.

Experimental workflow consists on preparing genomic DNA libraries and hybridizing them to capture arrays and then sequencing short targeted fragments. Each fragment usually is in the range of $300 - 500$ bps and specific adapters are added to their ends in order to allow their attachment and sequencing. Single DNA molecules separated in a solid support, called Flow Cell, are clonally amplified and in parallel sequenced.
Results of sequencing are generated by reading optical signals during iterative cycles of polymerase-mediated nucleotide extensions or, in one approach, through successive oligonucleotide ligations [18]. Data produced by a single run are more than 600 gigabases of nucleotide sequence, called *reads*, with a length ranging in $25 - 100$ bps. A such parallel sequencing process allows an overlapping of many random DNA fragments, therefore each nucleotide, in target region, will be read in many different reads, determining his depth coverage. Depending on the design selected, we can have sequencing runs targeted to single end or paired end reads. In paired end runs a fragment is first sequenced from one end to the other end for up to $25 - 100$ bases (depending on what decided by the the researcher), then another round of reading is performed in the opposite way. The paired end designs improves mapping quality during the process of alignment as the corresponding DNA fragment in the reference genome can identified more confidently.
There are different commercially available exome sequencing platforms that capture a slightly different set of target regions, being some also fo-

cused on regulatory regions as promoters and exon-intron junction region. For example, the three sequencing platforms from *Agilent, Illumina* and *Nimblegen* target 51 Mb, 62 Mb and 64 Mb respectively, with 29.45 Mb in common among all and 4.428 Mb of unique target regions. However, efficiency and overall coverage of strict coding regions has been shown to be highly concordant [17].

Whatever is the platform utilized, data process and raw data checks have to follow standard rules to assure quality of data. Specifically, once sequencing reads have been generated, they are aligned to the reference genome and assigned a probability for the matching in a specific genomic position. Base quality scores have to be recalibrated to correct for variation in quality with machine cycle and sequence context. Reads that are identical to others (duplicated reads) are marked and removed, as they are likely to be potential PCR artifacts [21]. At this point is important a careful quality assessment of sequencing run performance, through summary statistics of key parameters. More details are given in section 3.2.1. After mapping and quality process, reads are then analyzed to call genotypes at polymorphic sites, a process known as variant calling. Details are given in section 3.2.2. Variants are then filtered with a strategy that will depend on the hypothesized inheritance model, diseases penetrance and expected frequency of the causative variance, and prioritized according to their biological predicted function. This process of filtering has been very successfull for monogenic diseases, where only one, highly penetrant, deleterious variant is expected

to be the causal mutation. For complex diseases, where multiple variants can contribute to the clinical outcome, we are often unable to discriminate real disease-causing genetic variants from the broader background of rare variants present in all human genomes, which are not pathogenic for the disease under investigation. In such cases, more extensive genetic analyses or biological characterization of variants are necessary to pinpoint to the causative site.

1.2.2 Low-Pass DNA Sequencing of 2120 Sardinians

The huge data resources available from international large scale projects, such as *1000 Genomes projet*, allowed the discovery of a large number of low and rare novel variants in individuals from a broad set of populations, including admixed samples [14]. Furthermore, it has emerged that low-frequency variants show substantial geographic differentiation [14] and that founder isolated populations are likely to underrepresented in this large effort. For example, it was shown that, 1000G Consortium identified 50%, 98% and 99.7% of the SNPs with frequencies of 0.1%, 1.0% and 5.0% identified after sequencing 2500 samples from UK. By contrast, only 23.7%, 76.9% and 99.3% of SNPs in the same frequency range identified after sequencing 2000 Sardinian samples were also detected in the 1000 Genomes project. Therefore, despite this large international efforts, population-specific whole-genome sequencing studies are thus valuable,

in particular for the characterization of rare population-specific variants. Many research groups are undertaking whole-genome sequencing in their population of interest. In Sardinia, one large effort, carried out by the IRGB-CNR Institute and led by prof. Francesco Cucca in collaboration with prof. Goncalo Abecasis of the University of Michigan, is the Sardinia Medical Sequencing Project. It involves whole-genome sequencing of 2120 Sardinian samples at an average coverage of $4x$, whom 1122 are volunteers of the *SardiNIA project* [25] and others were enrolled in a Multiple Sclerosis (MS) and Type 1 Diabetes (T1D) case-control study [24]. The sequencing effort led to the discovery of $17,617122$ single-nucleotide polymorphisms (SNPs), of which 21.6% were not identified in any other population (based on dbSNP 142 and the Exome Aggregation Consortium)[3]. For the work I present in this PhD thesis, I was able to use this resource to flag and discard variants shared among affected family members but that were unlikely to be associated with the diseases, being rare in elsewhere in Europe but common in Sardinia, or absent elsewhere by present in other non-affected Sardinians.

[3] http://exac.broadinstitute.org/

Chapter 2

Genetics of Type 1 Diabetes

2.1 Many faces of diabetes

Type 1 diabetes (T1D) is a common, complex and autoimmune disease where genetic, epigenetic and environmental factors contribute to risk. It manifests itself through an autoimmune destruction of pancreatic β cells, resulting in a lack of production of insulin [28]. Data from large epidemiologic studies showed that over the last few decades, there has been an increase in the worldwide incidence of T1D by $2-5\%$ [29], suggesting the importance of environmental factors in the etiology of disease [34]. On the other hand, genetic predisposition is evident from the significant familial clustering. In fact, the average prevalence risk in children of an affected parent ranges from 2 to 8%, in dizygotic twin is about 8%, and in

monozygotic twins is as high as 50% (30% within 10 years of diagnosis of the first twin), numbers that are considerably higher than the 0.4% risk of the general population [30][31][33].

Type 1 Diabetes represents the most common type of diabetes in children and young adults and historically was known as *Juvenile Diabetes*; the typical adult form of diabetes is instead known as Type 2 Diabetes (T2D).

Diabetes tends to be divided into these two major categories, but other minor and monogenic forms exist, highlighting the wide disease's heterogeneity. For example, a common subgroup of T1D and T2D is *LADA (latent autoimmune diabetes*), which appears to be an admixture of the two forms of diabetes, sharing common symptoms and underlying genetic risk factors [36][39]. LADA is usually diagnosed over $30 - 35$ years of age and for at least the firsts six months is unlikely to need insulin treatment [38]. The diagnosis is mostly based on GAD antibody positivity, a parameter also used for T1D diagnosis [36], therefore misspecification is not infrequent [38][36].

There also exist monogenic forms of diabetes, as *MODY (Maturity-Onset-Diabetes of the Young)*, a subtype of familial diabetes characterized by early onset (usually before 25 years old in Caucasians) and specific autosomal dominant mutations in thirteen genes [36][40]. However, heterogeneity is present also in this class.

It is likely that many familial forms are clinically misclassified as type 1 or type 2 diabetes, as often the excess of affect members contrasts with the

expectations based on population incidence estimates. In depth studies of such families can help to understand the genetics of these peculiar disease manifestations, and will provide new insights into the pathogenesis of more common forms of diabetes.

2.2 From linkage studies to whole-exome sequencing in Type 1 Diabetes

Multiple linkage studies showed that the *HLA* locus, on chromosome $6p21.3$, represent the major risk factor for T1D, explaining up to $40-50\%$ of the familial clustering [32]. [1] Non-HLA loci, contributing to T1D, in particular *INS* gene, *CTLA*4, *PTPN*22 and *IL*2*RA*, have been associated through studies based on gene candidate approach.

Since 2007, several GWAS and meta-analysis have confirmed the role of such genes and up to date 57 independent T1D susceptibility loci outside the major HLA are established. Many of these loci contain genes relevant for the immune response and also expressed in pancreatic beta-cell, reinforcing the idea that immune system and beta-cell function are important for T1D pathogenesis [36][37][41][42]. Overall, the identified genetic factors explains a large fraction of the heritability (about 80% , of an overall estimate of 88% [43])[36].

[1] This strong association with HLA did not consistently replicate in T2D, confirming that T1D and T2D are two distinct diseases.

Although this is an exception compared to other complex human diseases, many familial T1D cases cannot be explained by their HLA status or higher genetic burden of risk-alleles at the 57 non-HLA loci. Therefore, genetic studies are now concentrating on families and applying novel technologies, as next-generation sequencing, to achieve a deep characterization of the genetic background. A successful effort has been reported in *Biason-Lauber et al.*, where authors studied a family with multiple individuals, of which four were affected by T1D, one by ulcerative colitis and someone had an unclear diabetes phenotype. By using a combined approach of different technologies −− microsatellite genotyping, targeted deep sequencing, exome-sequencing and Sanger sequencing of relevant candidate genes −− , they identified a mutation in the *SIRT1* gene carried only by the members affected by an autoimmune disorder [45], therefore suggesting a new mutation for a monogenic form of T1D [45].

2.3 Type 1 Diabetes in Sardinia

Epidemiological studies consistently show that Sardinia and Finland represent the two most high-risk areas for T1D worldwide [46]. Furthermore, a 5-fold increased prevalence of T1D has been observed in Sardinian MS patients, in the same individuals and/or in the same families. From the study of *Marrosu et al.* has emerged that associations of both T1D and MS with common variants in the HLA region explain only a small part of the

co-occurrence of these two autoimmunes diseases [44], suggesting the involvement of one or more damaging Sardinian-specific variants affecting the immune system.
Investigations in families with more individuals affected by T1D in multiple generations, and/or also in presence of double autoimmune disorders, could represent a strength point in genetic studies of T1D, particularly in a genetically isolated population as Sardinia, where T1D (and MS) are common and there is evidence of powerful founder effects [47].

Chapter 3

Study design and Methods

3.1 Subjects and study design

The family used in this study was initially enrolled on 1993 by the team led by Prof. Francesco Cucca. Fresh specimens of already enrolled volunteers and extension to the newly affected individuals was carried out from 2010. All members are of Sardinian origin and there is not consanguinity. A written informed consent was obtained from all participants (14 individuals). Among all volunteers, 10 were diagnosed Type 1 Diabetes trough autoantibodies (GAD, IA2 and IAA) tests and two also suffered of a second autoimmune disease (Celiac disease or Multiple Sclerosis). Age at disease onset is highly variable, from 6 to 62 years (average 18.12 years). The other volunteers were healthy individuals, although some of them were posi-

tive for autoantibodies tests. A graphical representation of the pedigree is showed in Figure 3.1.

The pattern of inheritance among the affected family members was indicative of an autosomal dominant mutation. A linkage analysis was previously performed with microsatellite genotyping. Here, we used whole exome sequencing and linkage analysis with the integration of SNPs derived from commercial whole-genome genotyping arrays and from exome sequencing. Further, a previous linkage analysis with microsatellite genotyping has been performed.

We sequenced the exome of four members from this family: three affected individuals across three different generations and one healthy, as showed in Figure 3.1. Exome sequencing was performed in a single experimental run with other 38 individuals (sixteen T1D affected parent-offspring pairs and 6 individuals affected by Multiple Sclerosis or Type 1 Diabetes from other Sardinian multigenerational families).

Furthremore, thirteen members of the family were extensively genotyped through Ilumina OmniExpress beadchips, a genotyping array targeting about $750K$ SNPs. The hypothesis of this study is to identify rare (MAF $< 1\%$) genetic variants with large effect size, therefore a joint linkage analysis with exome sequencing in this family seems optimal because the same mutation can be observed in many relatives and cosegregation with T1D can be tested.

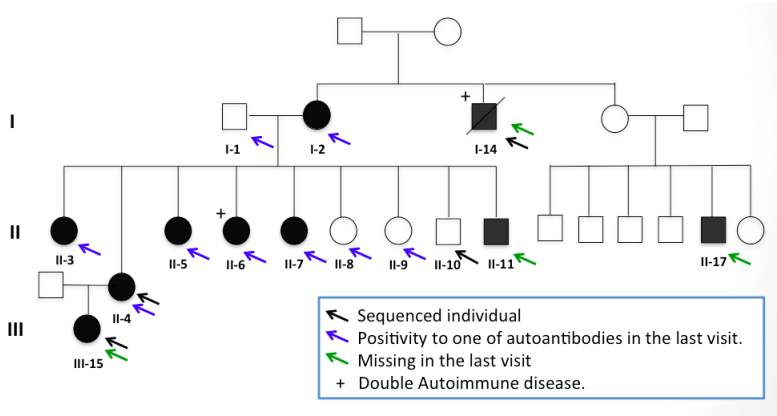


Figure 3.1: The pedigree of the family enrolled. Affection status of each member and individuals selected for sequenced are highlighted as described in the legend.

3.2 Whole exome sequencing

Whole exome of 42 individuals was captured by the Illumina TruSeq Exome Enrichment Kit, followed by sequencing using Illumina HiSeq 2000 sequencer. This Illumina kit provides an high uniform coverage across 62 Mb of exomic sequence, including 5′ UTR, 3′ UTR, microRNA, and other non-coding RNA, for a total of 20,794 genes. The experiment was designed to sequenced paired-end reads of 100-bp length, and each sample was run in different lanes to reach the targeted coverage (60x). The 42 samples were sequenced in 5 runs, with twelve subjects been replicated in two of these to improve their initial coverage, as showed in figure 3.2.

The reads were aligned to the hs37d5 reference assembly from the 1000

Samples	MeanDepth	Colonna1
	First run	Final Run
1	44.57	74.08
2	38.31	63.91
3	42.96	68.96
4	38.56	58.77
5	35.44	53.02
6	39.99	59.04
7	35.06	55.72
8	27.84	46.15
9	25.45	46.27
10	26.44	47.29
11	30.05	50.08
12	33.64	54.77

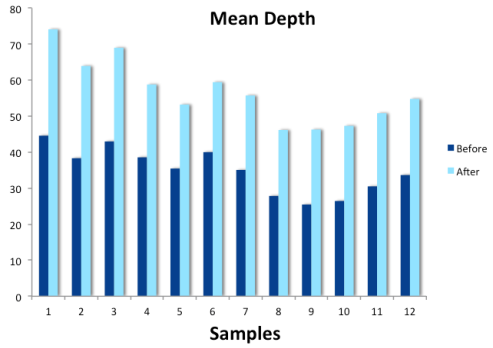


Figure 3.2: Differences in mean depth for 12 samples in first and in final run (a). Histograms of mean depth. (b).

Human Genome Project through Burrows-Wheeler Aligner (*BWA* [48]) version 0.6.2 and alignments for each sample were converted to a binary format for storing sequence data (BAM format), sorted by reference coordinate and indexed. A BAM file format (as well as in a SAM format - similar to bam, but in readable text format) is subdivided in two sections: header and alignment. The header section contains general information about the file, such as BAM file format version and sorting order of the alignments, which can be sorted by the reference coordinates, by query names, or unsorted. Other information include the sample identification code, the read group, the lane or the program used for alignement. The alignment section provides information for each sequence about mapping and quality with respect to the reference genome. By a *Cigar String* for

each sequence read, it is possible to know the number of bases that match or mismatch with the reference, how many are deleted or inserted.
At this point information for the same sample derived from different lanes was merged in an unique BAM file which was then processed with the *dedup* option in *bamUtil*[1] to determine duplicates. Then annotated with *MarkDuplicates* - an utility from the package *Picard-tools 1.81*[2] for marking PCR duplicates. In fact, duplicates may cause biases and alter variant calling results, because sequencing errors will be propagated in duplicates. After marking duplicates, the caller will be able to consider only one read among the duplicates and will more likely work in the right way.
After this step, reads were re-aligned if they map near known polymorphisms or insertions and deletions (according to the *GATK*[3] pipeline (version 2.7.4), then base qualities were recalibrated after all these processes. Finally, overlapping read pairs were cropped with the ClipOverlap option on the bamUtil executable. Each BAM file is then assessed for quality before being taken forward for the variant calling step, that we carried out with the tool *GATK HaplotypeCaller* to call indels and SNPs simultaneously. A schematic view of the described process is represented in Figure 3.3.

[1] http://genome.sph.umich.edu/wiki/BamUtil
[2] http://picard.sourceforge.net/
[3] https://www.broadinstitute.org/gatk/

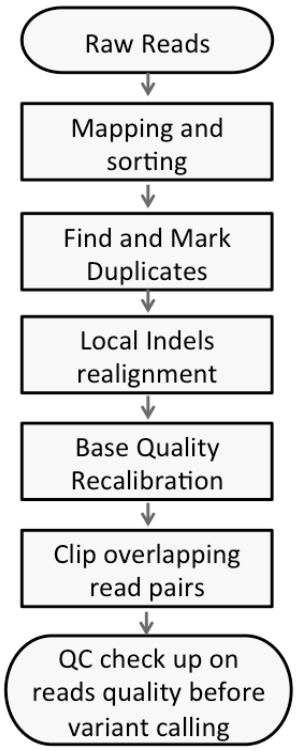


Figure 3.3: Workflow applied to NGS data processing

3.2.1 Quality Controls of reads

Data quality assessment represents a crucial step in next generation sequencing studies and it is much more complex in respect to traditional array platforms. Therefore, in order to diagnose sequencing run problems, a series of quality assessments was performed for all samples and

in all sequencing runs. Data quality is assessed through empirical and reported base quality score by comparing aligned bases to the reference genome. Empirical base scores are calculated in *Phred Score* Q, with $Q = -10 \cdot log_{10}(P_{err})$, where $P_{err} = Prob$(Error of called base). For example, Phred Score of 20 indicates that chances that this base is called incorrectly are 1% and a Pred Score of 30 correspond to an error probability of 0.1%, or, in other words, a base call accuracy of 99.9%.

In figure 3.4, empirical against reported Phred scores from two differents kits (SureSelect Target Enrichment System Kit and the newest TruSeq Exome Enrichment Kit) for five samples is shown. In the left panel, a clear deviation from the diagonal and overall lower quality is evident, indicating that the run was problematic. These samples (except one) have been sequenced a second time through the kit chosen for this study, and the empirical versus reported Phred score, represented in the right panel, shows that the experiment was now carried out properly. Other quality controls are showed in figures 3.6 and 3.7. For example, distribution of insert size shows that mean values range between 230 and 270 bp, which is close to the expected mean insert size for this Illumina kit (where mean expected insert size is 230 bp). Another factor that is important to consider in NGS experiments is the GC content, as it interferes with library amplification [49]. Therefore regions with extreme GC content may be under-represented, in accord with the figure 3.6, where low and high GC regions have low coverage, suggesting a higher error rate for these regions.

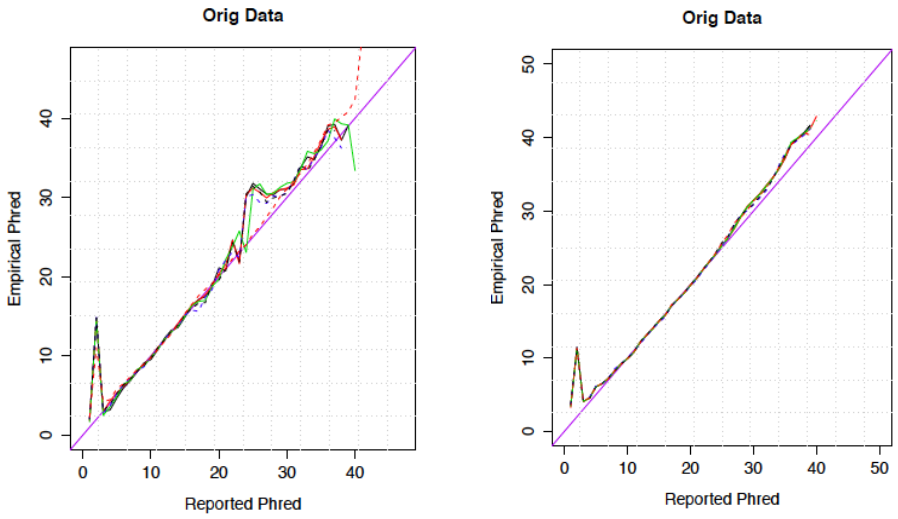


Figure 3.4: Empirical Phred score in a run showing deviation from Reported Phred (a) and in a run with a better trend (b)

Statistics as mean depth, empirical Q20 counts, number of mapped reads, percentage of duplicates as well as others are then compared among all samples in order to identify batches effects and other heterogeneities between samples. In figure 3.8 the number of total, mapped, paired, duplicated and failed reads for each sample in a specific run are plotted. These flag statistics have been useful not only to see differences across samples in the same run, but also in different sequencing runs.

After mapping and quality filters, a total of 112 million of reads have been mapped on target regions, with a mean depth of 57.19x and 98% with at

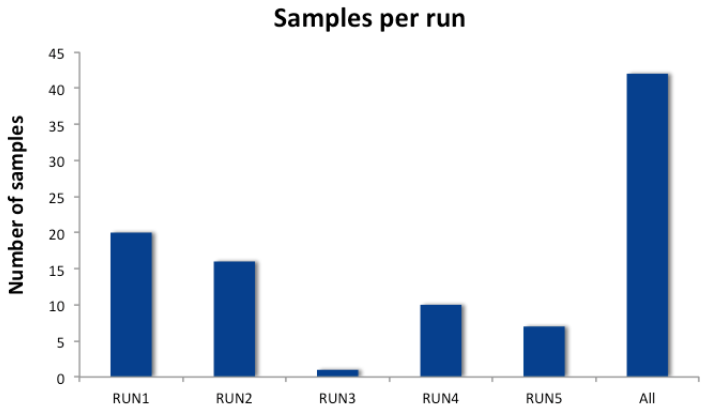


Figure 3.5: Number of samples for each run. Run1 and run2 have twelve subjects in common.

least a *Q-score* of 20.

3.2.2 Variant calling and analysis of variants

After quality checks, realignments and recalibration procedure, we started the process of variant calling. We used *HaplotypeCaller* of *Genome Analysis Toolkit (GATK)*, which calls SNPs and short indels simultaneously across all BAM files. Haplotype Caller employs a local de novo assembly algorithm that represents an agnostic approach with regards to variant type and divergence from any reference[22]. Furthermore, by jointly analyzing several samples, it extracts information from the other sequenced chromosomes,

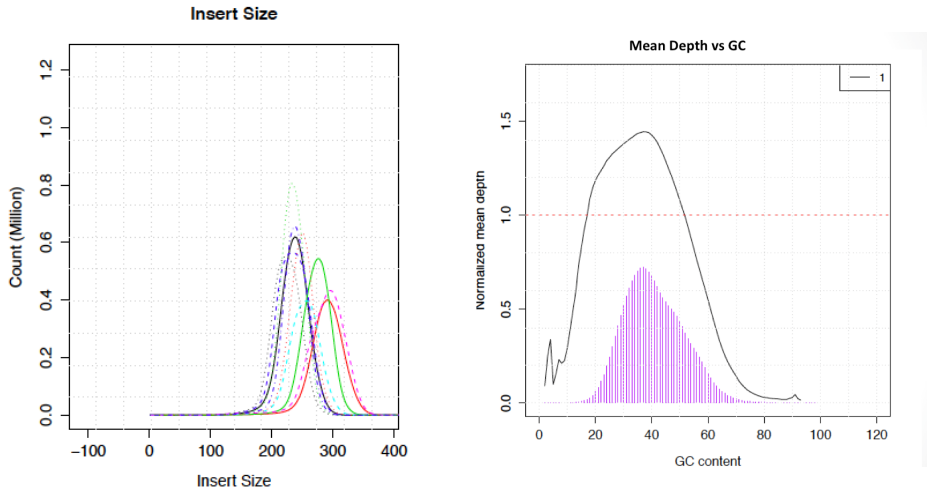


Figure 3.6: Insert size distribution in a run (a). Mean depth vs GC in a sample (b).

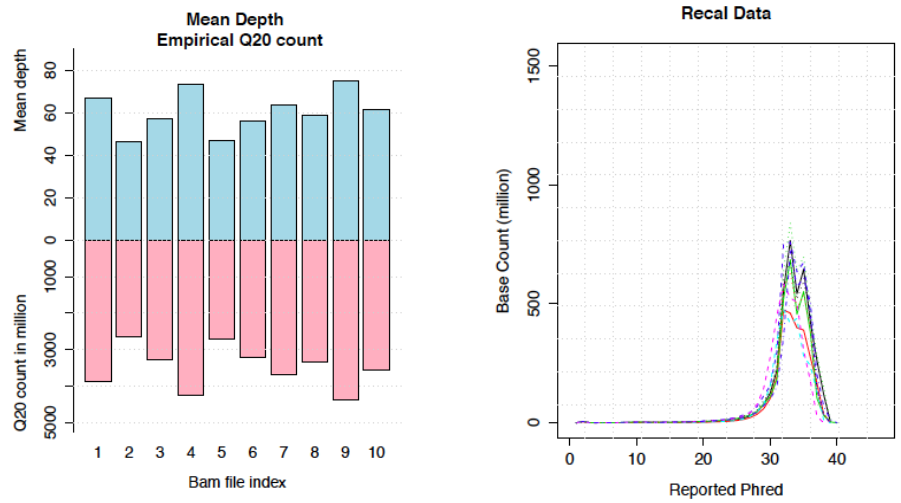
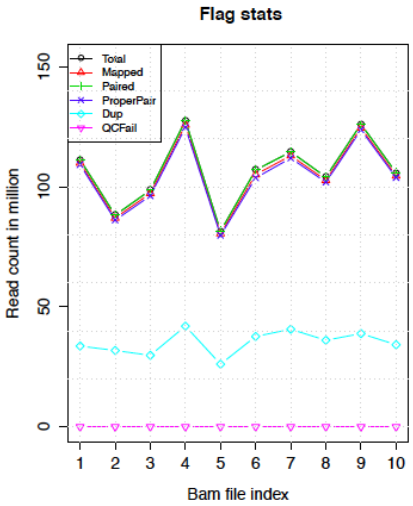


Figure 3.7: Mean depth and empirical Q20 count (a). Base count in milion vs reported Phred Score in a recalibrated run data (b).



All runs	Average Values
TotalReads (millions)	115.32
MappingRate (%)	96.67
ProperPaired (%)	94.82
Q20Bases (%)	98.05
MeanDepth	57.19
Duplicate Rate	35.34

Figure 3.8: Flag statistics in one of the analyzed runs (a). Average values of principal flagstats in all runs (b).

resulting in higher variants discovery rates. In addition to Haplotype Caller, in accord to Best Practices from GATK forum, a quantification of the goodness of called variants was performed through *Variant Quality score Recalibrator*, that assigns accurate confidence scores to each mutation and by statistical models is able to filter out false positive calls.

Variants are stored by the program in a VCF (Variant Calling Format) file. This is a text file with a specific format that is widely recognized. It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome. The header section includes information about format file, genome reference file and a key for each

annotation data. The section of variants has nine principal field:

- chromosome;
- physical position;
- rsID;
- reference allele and alternative allele(s);
- a quality score that depends on the caller;
- a binary filter (PASS or not PASS) if a variant quality score has been applied;
- an INFO field with some information that can be for example the number of alternative alleles in genotypes and alternative allele frequency, total number of alleles in called genotypes, total depth in this region and other annotations useful to define a variant quality score;
- a FORMAT field to define following information about genotypes in sample fields.

Subsequent columns are those corresponding to each individual genotype, written as 0/0, 0/1 or 1/1, where 0 represents the reference allele and 1 the alternative allele. Generally, after the genotypes there are allelic depth and a genotype likelihood, separated by a colon. A quick analysis of the VCF

file with *vcftool*[4] allows to have a general idea about the overall number of variants, divided by SNPs and indels, number of shared and private variants across all samples and number of transitions and transversions changes and their ratio. Graphics below show barplots for each of these numbers in all samples, highlighting similarity in their frequencies, except in the sample 26, which is the only individual sequenced with a different kit.

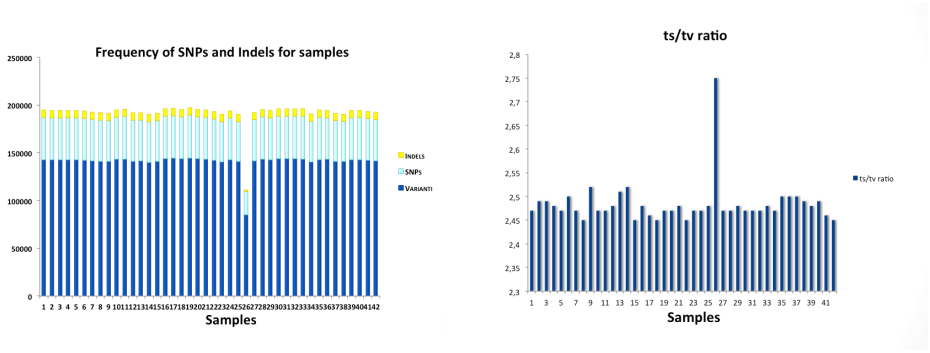


Figure 3.9: Barplots of overall number of SNPs and indels for sample (a). Transitions-transversions ratio for sample (b).

[4]http://sourceforge.net/projects/vcftools/

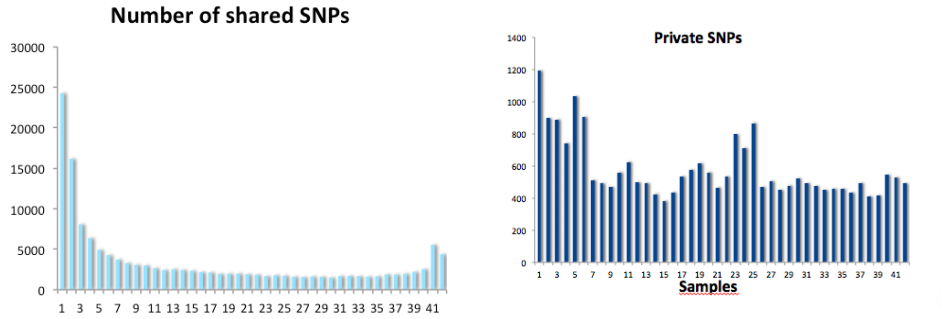


Figure 3.10: Barplots of shared SNPs across all samples(a). Barplot of private SNPs for sample (b).

3.2.3 Strategy of analysis

The hypothesis of this study is that one or more very rare mutations can be responsible for T1D within the family. The strategy is to focus on variants with a strong functional impact, highly penetrant and very rare in Sardinia and elsewhere. Therefore, we expect this variant(s) to be absent in the other sequenced individuals. We therefore proceeded by assessing the biological function of each variant and evaluating the frequency in several data sets, as described below.

The annotation of variants functionality has been performed with *Annovar tool*[5], which not only annotate single nucleotide variants and inser-

[5]http://www.openbioinformatics.org/annovar/

tions/deletions, but also reports functionally importance scores, as across-species convservation score. It also annotates the VCF INFO column if the variant is already reported in the 1000 Genomes Project, Exome Sequencing Project and dbSNP [51]. The variants found in the VCF but absent in these three datasets were considered Sardinian specific or very rare. In addition to these annotations, I considered also variants identified in the Sardinian whole-genome sequencing data reported in the Sardinian sequences reference panel [24]. We discarded all variants that were common in those four data sets (frequency> 1%), and retained those that were less frequent or for which frequency in the population was not reported (as per dbSNP). We then marked the remaining variants to note those that were present in the other Sardinian sequenced individuals (low-pass or exome-sequencing) that are not part of the T1D family under study.

The VCF analysis showed that the total number of variants discovered in the full exome-sequencing effort (42 individuals) is 150, 812, whom 134, 266 are SNPs and 16, 546 are indels. After the *Variant Quality Score Recalibrator* the total number of variants that passed our quality filters (PASS) is 141, 058 - of those 55, 037 are exonic variants, with different exonic functions, as showed in figure 3.11. Focusing only on the affected members of the family, the number of total PASS variants is 124, 047 and drops to about 51, 000 when considering only variants with an annotated minor al-

Exonic function	Variant Counts
Frameshift Del	303
Frameshift Ins	139
Frameshift sub	2
Deletions	460
Insertions	229
Substitutions	4
Stoploss	21
Stopgain	289
unknown	608
Synonymous	26037
Nonsynonymous	26945

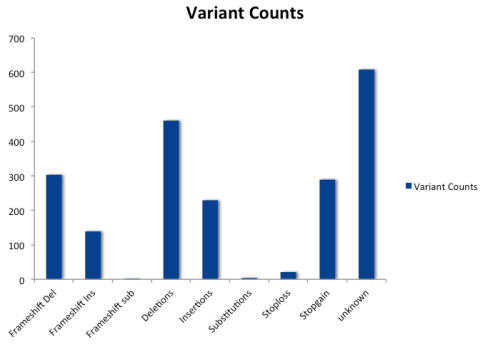


Figure 3.11: Counts of exonic variants (a). Barplot of exonic variants (b).

lele less than 1% or with unknown frequency. In Figure 3.12 we show the subsequent criteria used to prioritize variants, when applied to all variants detected or focusing only on specific biological types.

Despite the number of variants is drastically reduced following these filterings, it was still difficult to discriminate which were shared only by means of familial inheritance. We therefore considered linkage analysis as a useful integration to understand which variants fall in regions that are co-segregating with the disease.

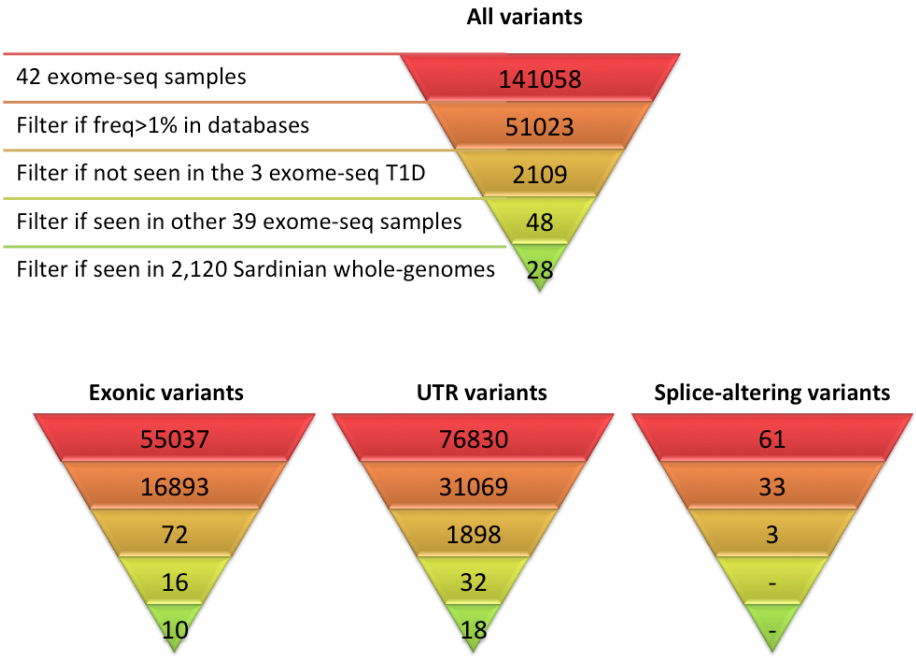


Figure 3.12: Counts of variants in whole exome sequencing analysis before and after frequency filtering, comparing all samples and the three affected individuals of the family.

3.3 Linkage Analysis

To further characterize exome-sequencing findings we carried out linkage analysis by incorporating genotypes from the *Illumina OmniExpress* array and carried out three parametric multipoint linkage analysis. Genotyping of the OmniExpress array was performed for all 13 members of the family

together with genotypes for other 2300 Sardinian individuals. Experiments were carried out in the *Lanusei UOS* of the *IRGB Institute* according to manufacturers protocols, and genotypes called with the *Illumina software GenomeStudio*. We performed standard per sample and per SNP quality checks considering the full set of OmniExpress arrays. Specifically, we required all samples to have a call rate > 98%, we removed SNPs with minor allele frequency < 1%, with genotyping call rate < 98%, with strong deviation from Hardy-Weinberg equilibrium ($p < 10^{-6}$), or that showed an excess of Mendelian inconsistencies. We then selected the 13 member of the T1D family and incorporated, for the 4 that were also sequenced, the genotype calls derived from exome-sequencing. The merged genotypes file was created with the option bmerge of *Plink tool*[6] [52] set in modality merge-mode 2, in order to use as a reference genotypes those from OE arrays and overwrite if missing in original. Excluding multiallelic SNPs and indels, the new merged file includes 794,575 variants, if which 95,756 are only from sequences and 112,019 were detected with both methods; the concordance rate was 90%. Linkage analysis have been performed with *Merlin*[7] [53]. According with the pattern of inheritance among affected members, as showed in the pedigree, we hypothesized for the linkage analysis a dominant model and set the disease allele frequency to 0.01%. To define population frequencies for linkage modeling, for each variant

[6]http://pngu.mgh.harvard.edu/purcell/plink/
[7]http://csg.sph.umich.edu/abecasis/Merlin/

given as input we estimated the frequency in the 2300 genotyped samples if was typed with the Omni array, and in the 42 sequenced samples if was derived with the exome-sequencing.

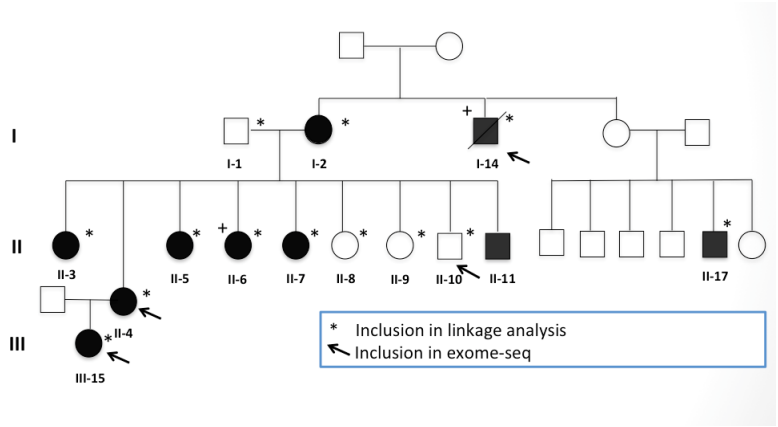


Figure 3.13: The patients'family tree showing selected individuals in the exome-seq and linkage analyses.

Moreover, since the program can only handle a limited number of family members, we performed three different multipoint linkage analysis, evaluating carefully the individuals to be included each time. The members of the first and third generation have always been considered, while there are differences in the second generation, resulting in the following three cases:

- *case 1*: only cases are included, except the individual II-17 (II-3, II-4, II-5, II-6, II-7);

- *case 2*: all cases (II-3, II-4, II-5, II-6, II-7, II-17);

- *case 3*: 4 cases and 2 controls (II-3, II-4, II-6, II-17, II-9, II-10). In this case, II-4 and II-6 were chosen because parent of III-15 and individual with dual pathology respectively, II-10 because sequenced in exome-sequencing and II-17 given that is the only in the second branch of the family.

3.3.1 Results

Results of the three parametric linkage analysis are shown in the table 3.1. All three cases converged on three regions: one on chromosome 2 (2q36.3-2q37.1), one on chromosome 13 (13q31.2-13q31.3) and one on chromosome 18 (18p11.31). However, only the linkage peak on chromosome 2 contains one of putatively causing exonic variants (non synonymous SNP in 2q37.1) detected in exome sequencing analysis.
Furthermore, this suggestive linkage signal had been seen in a previously genome linkage scan conducted for a part of the family (individuals I-1, I-2, I-14, II-3, II-4 II-5, II-6, II-7, II-8, II-9, II-10, II-11) using 593 microsatellites markers (398 purchased as commercial kits from *Applied Biosystems: ABI PRISM LINKAGE MAPPING SETS V2.5* and 195 additional markers, selected in the laboratory). The linkage analysis based on microsatellites markers, carried out for a dominant model with Merlin, indicates 2q37.1 as the strongest peak along with other suggestive signals on chromosomes

Chromosome	Cytoband	LOD Score Case 1	LOD Score Case 2	LOD Score Case 3
1	p36.12-p36.13	1.8	-	-
1	q25.3-q35.1	-	-	2.3
2	q36.3-q37.3	1.8	2.4	2.3
3	q26.31-q28	1.8	2.4	-
12	q14.3-q15	1.8	-	-
13	q31.2-q31.13	1.8	2.4	1.3
14	q31.3-q32.2	1.8	2.4	-
16	q22.1	-	-	2.0
17	q11.2-q21.32	-	-	2.3
18	p11.23-p11.31	1.8	2.3	2.3

Table 3.1: Results of parametric multipoint linkage analyses using joint exome-sequencing and array based genotypes, for case scenarios 1,2 and 3. Positions refers to the citogenetic location of the peak of LOD scores. LOD scores are reported if values were > 1.8 in at least one case.

14 (q31.3-q32.2) and 18, as showed in figure 3.14.

Our exome sequencing analyses pointed to variants that fall in one of the following linkage peaks that are seen in one or two case-scenarios: on chromosomes 1 (a one-base 3 UTR deletion q32.1), 3 (a 3 UTR SNP in 3q26.33), 12 (a 4-bases 3 UTR deletion in 12q15) and 14 (a 3 UTR SNP in

Figure 3.14: Results of a parametric multipoint linkage analysis on a part of the family using microsatellites data.

14q32.12) . Of note, the two SNP variants on chromosome 3 and 14 are not unique to the family, as they are present in the Sardinian reference panel with frequency of 0.17% and 0.33% respectively, whereas the 3 UTR deletions are private variants of this family. We then carried out validation trough Sanger sequencing of these five variants as well as other seven resulting only from exome sequencing analyses and that were predicted to have an high functional impact or that falled in genes with a role in known autoimmune pathways. The variants were sequenced in all 13 family members. The variants with the most consistent pattern of sharing between the affected individuals were three:

1. the non synonymous variant on 2q37.1, present in all affected members with the sole except of the more distant individual II-17;

2. the 1-base deletion on chromosome 1 at q32.1, present in all affected members with the sole exception of II-3;

3. the 4-bases deletion on 12q15 shared among all affected in the family, except II-11 and II-17.

The variant on 1q32.1 was absent in healthy members, while both other variants were present also in one healthy individual (II-8 and II-10 respectively). All other 9 mutations sequenced with Sanger did not show a pattern of genotypes that was consistent with the hypothesized disease's model.

These three variants therefore remain the most candidates and functional assessment of their biological role is now undergoing. The nonsynonymous mutation is not specific to Sardinians but is very rare elsewhere - it was described in the latest version of 1000 Genomes Project (Phase 3, November 2014) and in the Exome Sequencing Project 6500 with a frequency of 0.020% (1/5008 chromosomes) and 0.0077% (1/13006 chromosomes) respectively. SIFT and Polyphen2 predictions showed a damaging impact on protein function and a GERP score of 4.96 suggests an high index of conservation. If we limit to only individuals related to the more extended branch of the family, this variant shows the best co-segregation among affected members, although the penetrance is not complete. The other two deletions have not been previously described in public datasets and therefore at the moment are private mutations of this family. The

genes where those three variants fall are all potential candidates. The non-synonymous SNP falls within a gene that has been already suggested as a locus of T2D diabetes susceptibility, impaired glucose tolerance and insulin resistance. The other two deletions are located in the 3 UTR of genes with established roles in other autoimmune diseases like Celiac Disease and Multiple Sclerosis.

Chapter 4

Concluding Remarks

In this study I presented an application of exome sequencing data combined with a family-based linkage analysis on a family affected by type 1 diabetes.

The family presented here is unique for its number of affected individuals with T1D in three generations and co-occurrence of autoimmune diseases in two family members. The sequencing of 4 individuals, of whom 3 were affected and selected from different generations, produced a large number of potentially interesting variants that was then reduced to about 2000 variants after the application of several filters. Despite the considerable reduction of the initial number of variants, the set of candidate variants to be tested for validation was still too wide. We therefore applied a combined strategy that incorporates exome sequencing with array genotyping

data for the full family into a linkage analysis to evaluate co-segregation of these candidate variants with the disease. With this approach, it was possible to considerably reduce the number of variants to be validated by Sanger sequencing and consequently narrow the list of candidates to three mutations that are very rare in the general population (frequency < 0.01%). The variants fall in genes with a potential involvement in T1D, supporting that the results are genuine.

None of three mutation show a complete Mendelian co-segregation with the disease, however heterogeneity is expected in a complex disease such as T1D and some of the affected individuals can be a sporadic T1D patient. Moreover, even at the phenotypic level, the family is quite heterogeneous, not only in the age of onset, but also in the ambiguity concerning the analysis of autoantibodies, for which some of the healthy individuals were positive (including the father of the generation I). It is possible that this phenotypic diversity can result in heterogeneity at the genetic level. Finally, in the absence of functional data in support of a biological role in diabetes for these variants, it cannot be excluded that none of them is causative. Exome sequencing is indeed not an exhaustive assessment of the genomic variation therefore other classes of variants not included in this analysis may be causing the disease. For example, the approach may miss mutations in exonic regions that were not targeted by the exome-sequencing capture method utilized. Furthermore, structural exonic or genomic variants, as copy number variants, were not considered at all. Therefore, further ge-

netic characterization, including deep sequencing of key members, could reveal the presence of other variants causing diabetes in this family.

Bibliography

[1] Pritchard, J. K., Cox, N. J. *The allelic architecture of human disease genes: common diseasecommon variant... or not?* Hum. Mol. Genet. 11, 24172423 (2002).

[2] Maher B. *Personal genomes: The case of the missing heritability.* Nature. 2008;456:1821.

[3] Li, Yun et al.*Genotype Imputation.* Annual review of genomics and human genetics 10 (2009): 387-406

[4] Do, Ron, Sekar Kathiresan, and Gonalo R. Abecasis. *Exome Sequencing and Complex Disease:Practical Aspects of Rare Variant Association Studies.* Human Molecular Genetics (2012):R1-R9.

[5] Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A. and Shendure, J. *Exome sequencing as a tool for Mendelian disease gene discovery.* Nat. Rev. Genet. 12, 745-755

[6] Antonarakis, S. E., Beckmann, J. S. *Mendelian disorders deserve more attention.* Nat. Rev. Genet. 7, 277-282

[7] Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., Neale, B.M., McLaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L. et al. *Exome sequencing and the genetic basis of complex traits.* Nat. Genet. 44, 623-630

[8] Eggers S, Smith KR, Bahlo M, Looijenga LH, Drop SL, Juniarto ZA, Harley VR, Koopman P, Faradz SM, Sinclair AH. *Whole exome sequencing combined with linkage analysis identifies a novel 3 bp deletion in NR5A1.* Eur J Hum Genet. 2014 doi:10.1038/ejhg.2014.130

[9] Biason-Lauber, Anna et al. *Identification of a SIRT1 Mutation in a Family with Type 1 Diabetes.* Cell metabolism 17.3 (2013): 448-455

[10] Tanaka D, Nagashima K, Sasaki M, Funakoshi S, Kondo Y, Yasuda K, Koizumi A, Inagaki N. *Exome sequencing identifies a new candidate mutation for susceptibility to diabetes in a family with highly aggregated type 2 diabetes.* Mol Genet Metab. 2013;109(1):112-117.

[11] Dyment, David A. et al. *Exome Sequencing Identifies a Novel Multiple Sclerosis Susceptibility Variant in the TYK2 Gene.* Neurology 2012;79(5):406-411

[12] Manolio T.A., Collins F.S., Cox N.J., Goldstein D.B., Hindorff L.A., Hunter D.J., McCarthy M.I., Ramos E.M., Cardon L.R.,

Chakravarti A. *Finding the missing heritability of complex diseases.* Nature. 2009;461:747753.

[13] Schork, Nicholas J. et al. *Common Vs. Rare Allele Hypotheses for Complex Diseases.* Current opinion in genetics and development 19.3 (2009): 212219.

[14] The 1000 Genomes Project Consortium. *An Integrated Map of Genetic Variation from 1,092 Human Genomes.* Nature 491.7422 (2012): 5665.

[15] Tennessen, Jacob A. et al. *Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes.* Science (New York, N.Y.) 337.6090 (2012): 6469.

[16] Marth, Gabor T et al., the 1000 Genomes Project. *The Functional Spectrum of Low-Frequency Coding Variation.* Genome Biology 12.9 (2011): R84.

[17] Clark, Michael J et al. *Performance Comparison of Exome DNA Sequencing Technologies.* Nature biotechnology 29.10 (2011): 908914.

[18] Lin, Xi et al. *Applications of Targeted Gene Capture and Next-Generation Sequencing Technologies in Studies of Human Deafness and Other Genetic Disabilities.* Hearing research 288.0 (2012): 10.1016/j.heares.2012.01.004.

[19] Clark, Michael J et al. *Performance Comparison of Exome DNA Sequencing Technologies.* Nature biotechnology 29.10 (2011): 908914.

[20] Li, Dalin et al. *Using Extreme Phenotype Sampling to Identify the Rare Causal Variants of Quantitative Traits in Association Studies.* Genetic epidemiology 35.8 (2011): 790799.

[21] Ng, Sarah B. et al. *Targeted Capture and Massively Parallel Sequencing of Twelve Human Exomes.* Nature 461.7261 (2009): 272276.

[22] Iqbal, Zamin et al. *De Novo Assembly and Genotyping of Variants Using Colored de Bruijn Graphs.* Nature genetics 44.2 (2012): 226232.

[23] MacArthur, D. G. et al. *Guidelines for Investigating Causality of Sequence Variants in Human Disease.* Nature 508.7497 (2014): 469476.

[24] Pistis G, Porcu E, Vrieze S et al. *Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs* Eur J Hum Genet (2014)

[25] Pilia G, Chen WM, Scuteri A et al. *Heritability of cardiovascular and personality traits in 6,148 Sardinians.* PLoS Genet 2006; 2: e132.

[26] De La Vega FM, Bustamante CD, Leal SM. *Genome-wide association mapping and rare alleles: from population genomic to personalized medicine.* Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing 2011:74-75.

[27] Turkmen, Asuman S., and Shili Lin. *Blocking Approach for Identification of Rare Variants in Family-Based Association Studies.* Ed. Zhaoxia Yu. PLoS ONE 9.1 (2014): e86126.

[28] Borg WP, Sherwin RS. *Classification of diabetes mellitus.* Intern Med. 2000; 45: 279-95

[29] Maahs, David M et al. *Chapter 1: Epidemiology of Type 1 Diabetes.* Endocrinology and metabolism clinics of North America 39.3 (2010): 481-497.

[30] Cernea S, Dobreanu M, Raz I. *Prevention of type 1 diabetes: today and tomorrow.* Metab Res Rev. 2010;26(8): 602-605.

[31] Huber A, Menconi F, Corathers S, Jacobson EM, Tomer Y. *Joint genetic susceptibility to type 1 diabetes and autoimmune thyroiditis: from epidemiology to mechanisms.* Endocr Rev. 2008;29(6): 697-725.

[32] Risch N. *Assessing the role of HLA-linked and unlinked determinants of disease.* Am J Hum Genet 1987; 40: 1-14

[33] Kyvik, K.O., Green, A., Beck-Nielsen, H. *Concordance rates of insulin dependent diabetes mellitus: a population based study of young Danish twins.* BMJ 311, 913-917.

[34] Gale, E.A. *The rise of childhood type 1 diabetes in the 20th century.* Diabetes 51, 33533361.

[35] Pandey JP, Zamani M, Cassiman JJ. *Epistatic effects of genes encoding tumor necrosis factor-alpha, immunoglobulin allotypes, and HLA antigens on susceptibility to non-insulin-dependent (type 2) diabetes mellitus.* Immunogenetics 1999; 49: 860864.

[36] Groop L, Pociot F *Genetics of diabetes - Are we missing the genes or the disease?* Mol Cell Endocrinol 382: 726739.

[37] Bergholdt, R, 2009. *Understanding type 1 diabetes genetics approaches for identification of susceptibility genes in multi-factorial diseases.* Dan. Med. Bull. 56, 139.

[38] Groop, L., Tuomi, T., Rowley, M., Zimmet, P., Mackay, I.R., 2006*Latent autoimmune diabetes in adults (LADA) more than a name.* Diabetologia 49, 19961998.

[39] Hernandez, Marta et al., on behalf of the Action LADA consortium. *Insulin Secretion in Patients with Latent Autoimmune Diabetes (LADA): Half Way Between Type 1 and Type 2 Diabetes: Action LADA 9.* BMC Endocrine Disorders 15.1 (2015)

[40] Bonnefond, Amlie et al. *Whole-Exome Sequencing and High Throughput Genotyping Identified KCNJ11 as the Thirteenth MODY Gene.* Ed. Klaus Brusgaard. PLoS ONE 7.6 (2012)

[41] Burren, Oliver S. et al. *T1DBase: Update 2011, Organization and Presentation of Large-Scale Data Sets for Type 1 Diabetes Research.* Nucleic Acids Research 39.Database issue (2011)

[42] Onengut-Gumuscu S1, Chen WM2 et al. *Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers.* Nat Genet. 2015 Mar 9. doi: 10.1038/ng.3245.

[43] Valma Hyttinen, Jaakko Kaprio, Leena Kinnunen, Markku Koskenvuo, and Jaakko Tuomilehto *Genetic Liability of Type 1 Diabetes and the Onset Age Among 22,650 Young Finnish Twin Pairs: A Nationwide Follow-Up Study* Diabetes April 2003 52:4 1052-1055.

[44] Marrosu MG, Motzo C, Murru R, Lampis R, Costa G, Zavattari P, Contu D, Fadda E, Cocco E, Cucca F.*The co-inheritance of Type 1 Diabetes and Multiple Sclerosis in Sardinia cannot be explained by genotype variation in the HLA region alone.* Hum Mol Genet 2004.

[45] Biason-Lauber, Anna et al.*Identification of a SIRT1 Mutation in a Family with Type 1 Diabetes.* Cell metabolism 17.3 (2013): 448455.

[46] Songini, Marco, and Cira Lombardo. *The Sardinian Way to Type 1 Diabetes.* Journal of Diabetes Science and Technology 4.5 (2010): 12481255.

[47] Pitzalis, Maristella et al. *Genetic Loci Linked to Type 1 Diabetes and Multiple Sclerosis Families in Sardinia.* BMC Medical Genetics 9 (2008)

[48] Li, Heng, and Richard Durbin. *Fast and Accurate Long-Read Alignment with BurrowsWheeler Transform.* Bioinformatics 26.5 (2010): 589595.

[49] Aird, Daniel et al. *Analyzing and Minimizing PCR Amplification Bias in Illumina Sequencing Libraries.* Genome Biology 12.2 (2011).

[50] Li, Bingshan et al. *QPLOT: A Quality Assessment Tool for Next Generation Sequencing Data.* BioMed Research International 2013 (2013).

[51] Wang, Kai, Mingyao Li, and Hakon Hakonarson. *ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data.* Nucleic Acids Research 38.16 (2010): e164.

[52] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC *PLINK: a toolset for whole-genome association and population-based linkage analysis.* American Journal of Human Genetics, 81(2007).

[53] Abecasis GR, Cherny SS, Cookson WO and Cardon LR. *Merlin-rapid analysis of dense genetic maps using sparse gene flow trees.* Nat Genet 30:97-101 (2002).